

На правах рукописи

Сергейко Анастасия Павловна

**КОМПЬЮТЕРНОЕ КОНСТРУИРОВАНИЕ ФЕРМЕНТАТИВНЫХ СИСТЕМ,
ОБЕСПЕЧИВАЮЩИХ СИНТЕЗ МАКРОЛИДОВ С ЗАДАННЫМИ
СВОЙСТВАМИ**

03.00.28 - биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Москва – 2008

Работа выполнена в Государственном Учреждении Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича Российской академии медицинских наук

Научный руководитель:

доктор биологических наук,
профессор

Поройков Владимир Васильевич

Официальные оппоненты:

доктор физико-математических наук,
профессор

Туманян Владимир Гаевич

доктор биологических наук,
профессор

Гельфанд Михаил Сергеевич

Ведущая организация:

Федеральное государственное унитарное предприятие Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов

Защита состоится 24 апреля 2008 года в 11:00 часов на заседании Диссертационного совета Д 001.010.01 при ГУ НИИ БМХ РАМН по адресу: 119121, Москва, ул. Погодинская, д. 10.

С диссертацией можно ознакомиться в библиотеке ГУ НИИ БМХ РАМН.

Автореферат разослан « 21 » марта 2008 г.

Ученый секретарь Диссертационного совета
кандидат химических наук



Е.А. Карпова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Среди применяемых в настоящее время лекарственных средств большую группу составляют природные соединения и их производные. Примерно 50% принципиально новых препаратов (NCE), открытых с 2000 по 2006 годы, являются природными соединениями [Newman D.J., 2007].

Макролиды – это одна из групп природных соединений, общей структурной особенностью которых является наличие макролактонового кольца. К группе макролидов относятся антибиотики, противогрибковые, противопаразитарные и противоопухолевые средства, иммуносупрессанты, и др. [Omura S., 2002]. Благодаря структурному разнообразию и многообразию проявляемых биологических эффектов, макролиды являются перспективной группой веществ для поиска среди них новых биологически активных соединений. Особенно актуален поиск новых антибиотиков в связи с развитием резистентности бактерий к уже существующим лекарствам.

В настоящее время поиск новых биологически активных макролидов осуществляется тремя путями: 1) скрининг биологических образцов; 2) химический синтез/модификация известных соединений; 3) модификация бактерий-продуцентов методами геной инженерии, приводящая к синтезу ими новых веществ. Последний подход активно развивается в последние годы и получил название «комбинаторный биосинтез» [Khosla C., 2006].

Разработка и применение методов комбинаторного биосинтеза стали возможны благодаря расшифровке механизма синтеза макролидов. Макролактоновое кольцо макролидов синтезируется из небольших углеродных фрагментов посредством повторяющихся реакций конденсации, подобно жирным кислотам. Этот процесс катализируется модульными поликетидсинтазами (поликетидсинтазами типа I). Модульные поликетидсинтазы (PKS) являются крупными ферментативными системами, состоящими из нескольких белков. Каждый белок имеет доменное строение, и соответственно, несколько центров, обладающих разными каталитическими активностями. Группа доменов, отвечающая за один цикл конденсации, образует «модуль» [Donadio S., 1991]. Модуль состоит минимум из 3-х доменов: кетосинтазного (KS), ацилтрансферазного (AT) и ацилпереносящего (ACP). Помимо этих основных доменов, модуль PKS может содержать дополнительные домены: кеторедуктазный (KR), дегидратазный (DH), енол-редуктазный (ER). Наличие или отсутствие этих доменов в модуле определяет степень восстановления той части, которая добавлена предшествующим модулем в поликетидную цепь. В процессе биосинтеза происходит последовательное удлинение строящейся поликетидной цепи на два атома углерода каждым модулем. Когда поликетидная цепь полностью синтезирована, она отделяется от PKS с помощью тиоэстеразного домена (TE), затем происходит ее циклизация “хвост

к голове”. Образовавшееся макролактоновое кольцо обычно подвергается пост-поликетидным модификациям (гидроксилирование, гликозилирование, ацилирование и др.), в результате чего окончательно определяются структура и биологические функции синтезируемой молекулы. Методами комбинаторного биосинтеза можно изменять количество и набор доменов и модулей в PKS, и, таким образом, изменять структуру синтезируемого PKS макролида.

Теоретически методами комбинаторного биосинтеза можно получить огромное количество новых макролидов [Gonzalez-Lergier J., 2005], однако провести генно-инженерные эксперименты по получению такого большого числа микроорганизмов-продуцентов невозможно. Также заранее неизвестно, будут ли синтезируемые ими вещества проявлять требуемую биологическую активность. Таким образом, является актуальной разработка компьютерного метода, позволяющего отбирать среди множества возможных вариантов макролидов структуры с требуемыми свойствами и определять состав необходимых для их биосинтеза ферментативных систем.

В настоящее время в литературе описано две попытки использования компьютерных методов для рационального поиска новых соединений из группы макролидов [Siani M.A. et al., 2000; Adalsteinsson H. et al, 2000]. Однако в этих работах генерация структур ограничивалась аналогами определенных макролидов, а в основе алгоритма отбора лежал механизм действия базового вещества. Общая задача генерации разнообразных структур макролидов в соответствии их метаболическими путями и отбора среди них перспективных соединений на основе прогноза совокупности видов их биологической активности в настоящее время не решена.

Цель работы. Разработка метода конструирования *in silico* ферментативных систем синтеза макролидов, генерации соответствующих комбинаторных библиотек вторичных метаболитов и отбора тех вариантов ферментативных систем, которые обеспечивают получение веществ с заданным профилем биологической активности.

Задачи исследования:

1. Создание компьютерной программы для генерации виртуальных библиотек структурных формул макролидов и описаний синтезирующих их ферментативных систем.
2. Создание базы данных, содержащей информацию о структуре и биологической активности макролидов.
3. Оценка применимости используемого метода прогноза биологической активности для предсказания биологической активности макролидов.
4. Апробация предложенного метода на примере генерации и скрининга виртуальной библиотеки аналогов эритромицина.

Научная новизна. Предложен оригинальный подход к конструированию доменного состава ферментативных систем, обеспечивающих синтез новых макролидов с требуемым профилем биологической активности.

Впервые созданы компьютерная программа Biogenerator, позволяющая генерировать виртуальные библиотеки структур макролидов и синтезирующих их ферментативных систем, и база данных, содержащая информацию о структуре и видах биологической активности более 1000 макролидов.

Практическая значимость. Предложенный метод обеспечивает возможность конструирования новых биологически активных структур макролидов и необходимых для их биосинтеза наборов доменов в модульных PKS. Сконструированные ферментативные системы могут быть реализованы в бактериях-продуцентах с помощью методов генной инженерии. Описанный метод дает возможность рационального планирования генно-инженерных экспериментов по модификации бактерий-продуцентов макролидов. Применение описанного подхода позволит значительно сократить число экспериментов по получению структур макролидов и тестированию их биологической активности, и, следовательно, существенно снизить временные и финансовые затраты при поиске новых лекарственных препаратов группы макролидов.

Апробация работы. Основные положения диссертационной работы были доложены на Московской международной конференции «Биотехнология и медицина» (Москва, 2006 г.), 14-й Международной конференции и дискуссионном научном клубе «Новые информационные технологии в медицине, биологии, фармакологии и экологии» (Словакия, Низкие Татры, 2006 г.), IV Московском международном конгрессе «Биотехнология: состояние и перспективы развития» (Москва, 2007 г.), научной конференции ГУ НИИ БМХ РАМН (Москва, 2007 г.), 4-ом Международном симпозиуме «Компьютерные методы в токсикологии и фармакологии, включающие Интернет-ресурсы (СМТРИ-2007)» (Москва, 2007 г.).

Публикации. По материалам диссертации опубликовано 6 печатных работ (3 статьи, 2 из них в рецензируемых журналах, 1 – в сборнике трудов конференции; 3 – материалы докладов на конференциях). Получено 2 свидетельства Роспатента об официальной регистрации компьютерных программ.

Объем и структура диссертации. Диссертационная работа изложена на 108 страницах, содержит 20 рисунков и 8 таблиц. Работа состоит из введения, обзора литературы, описания материалов и методов исследования, собственных результатов, их обсуждения, выводов, списка цитированной литературы, включающего 120 источников, приложения.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Предлагаемый нами подход состоит в генерации *in silico* структур макролидов в соответствии с механизмом их биосинтеза и последующем анализе полученных виртуальных библиотек на основе компьютерного предсказания биологической активности. Такой анализ позволяет предварительно отобрать наиболее перспективные вещества, согласно прогнозу обладающие требуемыми фармакодинамическими свойствами. Для структур макролидов определяется количество модулей, их последовательность и набор доменов в каждом модуле PKS типа I, которые обеспечивают синтез отобранных соединений, то есть конструируется ферментативная система, необходимая для биосинтеза определенной структуры.

Для создания базы данных по структурам макролидов и проявляемым ими видам биологической активности были использованы данные из следующих источников: Discovery Gate, PubChem, PubMed, ChemIdPlus, Heterocycles Database, Dictionary of Natural Products. Макролидами считали те вещества, для которых это было указано в соответствующих полях баз данных или в публикациях.

Biogenerator. Для генерации виртуальных библиотек ферментативных систем и синтезируемых ими структур макролидов нами разработана программа Biogenerator (см. «Результаты и обсуждение»). Расчетная часть программы написана на языке Си++, графическая оболочка – на языке скриптов Tcl/Tk.

PASS. Для прогноза биологической активности соединений использовалась программа PASS версии 2006 (Prediction of Activity Spectra for Substances - прогноз спектров биологической активности органических соединений) [Филимонов Д.А., Поройков В.В., 2006]. Под спектром биологической активности понимается весь комплекс биологических эффектов, которые вещество способно вызывать при определенных условиях взаимодействия с биологическими объектами. Биологическая активность представлена в PASS качественно (наличие/отсутствие). Программа PASS версии 2006 позволяет по структурной формуле вещества оценивать вероятность проявления (P_a) или отсутствия (P_i) около 2800 видов биологической активности.

В программе PASS в качестве основы для описания структуры органических соединений используется структурная формула. Для описания структурной формулы используются «MNA-дескрипторы» [Filimonov D.A., 1999]. Для прогноза в PASS используется SAR base, которая создается на основе анализа обучающей выборки, содержащей структурные формулы и известные из эксперимента спектры активности органических соединений. В PASS версии 2006 SAR base содержит данные о 68038 лекарствах и биологически активных веществах. Словарь MNA-дескрипторов включает 47858 уникальных дескрипторов 1-го и 2-го уровней.

Оценка точности прогноза

Для оценки точности прогноза использовали инвариантный критерий точности прогноза – IAP (Invariant Accuracy of Prediction). Для каждого вида активности IAP подсчитывается по формуле:

$$IAP = \frac{N(Pa(s_a) > Pa(s_i))}{Na \cdot Ni},$$

где $N(Pa(s_a) > Pa(s_i))$ – число случаев, когда Pa для активных соединений превышает Pa для неактивных соединений; Na и Ni – число активных и неактивных соединений в выборке.

Средняя точность прогноза PASS версии 2006 составляет около 93% при скользящем контроле с исключением по одному.

PharmaExpert. Для анализа результатов прогноза PASS использовалась экспертная система PharmaExpert [Poroikov V. et al., 2004], описывающая взаимосвязи между фармакологическими эффектами и механизмами действия физиологически активных веществ.

ISIS/Base. Для создания базы данных структур макролидов, а также визуализации виртуальных библиотек макролидов и синтезирующих их ферментативных систем, была использована СУБД ISIS/Base (www.mdl.com).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Компьютерная программа Biogenerator для генерации виртуальных библиотек макролидов и описаний синтезирующих их ферментативных систем

Созданная нами программа Biogenerator позволяет моделировать комбинаторные перестановки доменов и модулей в поликетидсинтазах типа I и, в соответствии с этим, генерировать виртуальные библиотеки ферментативных систем и синтезируемых ими макролидов. Результаты генерации сохраняются в SDF-файлах, которые содержат структуры молекул и описания соответствующих им PKS. Описание PKS включает порядок следования и доменный состав модулей в гипотетическом ферментативном комплексе.

Входные параметры программы Biogenerator:

- I а) Тип стартового «строительного блока» (ацетат или пропионат).
- б) Размер макролактонового кольца молекулы макролида (6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38 атомов), определяющий число модулей в гипотетической PKS. Число модулей рассчитывается следующим образом:
 $M = N/2 - 2$, где M – число модулей; N – число атомов в кольце.

II Место и тип пост-поликетидных модификаций макролактонового кольца, задаются с помощью редактора параметров генерации. На основании своих представлений о наличии и специфичности действия соответствующих ферментов в данной биосинтетической системе, пользователь (специалист в области геной инженерии) может устанавливать различные типы пост-поликетидных модификаций в определенных позициях макролактонового кольца. Могут быть заданы следующие типы модификаций (рис. 1):

Гликозилирование. Сайтами гликозилирования являются предполагаемые гидроксильные группы, которые появляются в результате функционирования дополнительного домена KR.

Гидроксилирование. Сайты пост-поликетидного гидроксилирования могут быть установлены в определенных позициях макролактонового кольца, которые не перекрываются с теми, где гидроксильная группа может появиться при восстановлении β -кетогруппы в результате работы домена KR.

Окисление метильной группы. Для полиеновых макролидов, размер макролактонового кольца которых равен или больше 22 атомов.

Наличие/отсутствие полукетала. Полукеталь может появляться при взаимодействии гидроксильной группы и кетогруппы, расположенных через три атома углерода друг от друга в макролактоновом кольце. Задание полукетала накладывает ограничения на доменный состав двух модулей.

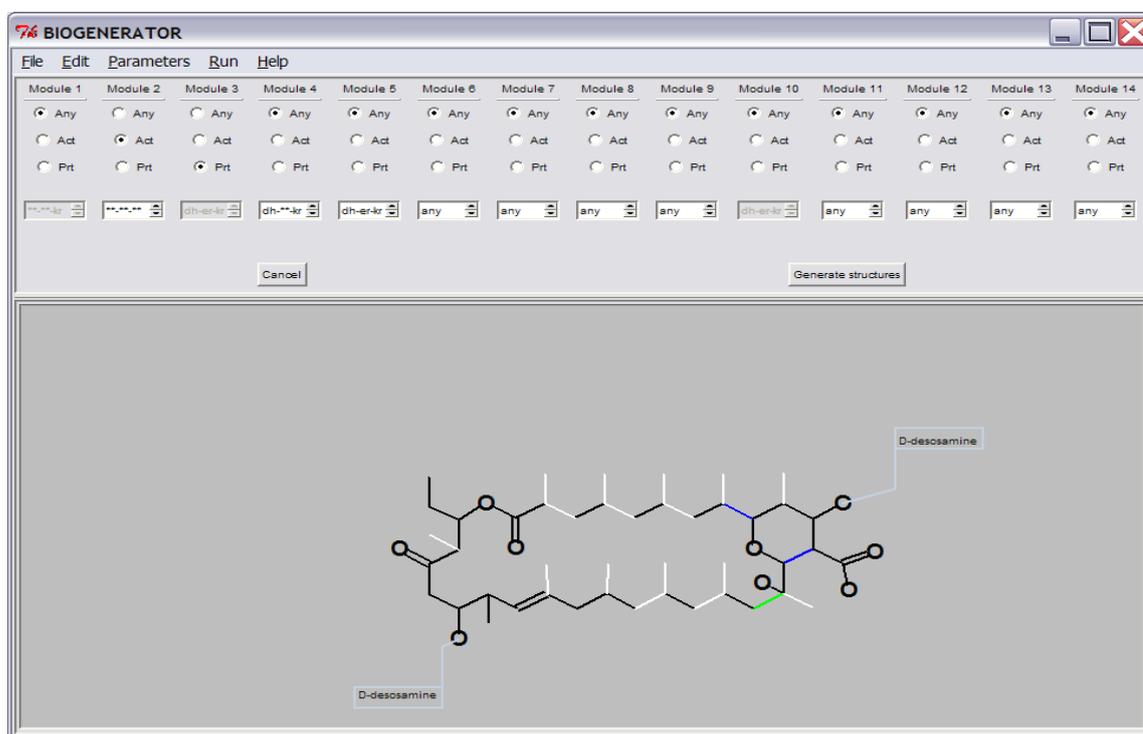


Рис. 1. Задание доменного состава отдельных модулей. В верхней части окна показаны модули и набор доменов в них, в нижней – соответствующее этому набору доменов макролактоновое кольцо. Показанные белым цветом связи в макролактоновом кольце изменяются в процессе генерации структур, черным – не изменяются.

После того, как выбраны параметры **I** и **II**, пользователь может также задать, какие домены не будут изменяться в процессе генерации. При задании доменного состава отдельных модулей (**III**) можно выбрать субстратную специфичность домена АТ и набор дополнительных доменов для каждого модуля (рис. 1). Фиксированному доменному составу всех модулей PKS соответствует определенная химическая структура, которая будет синтезироваться данной ферментативной системой.

После установления обязательных (**I**, **II**) и дополнительных (**III**) параметров начинается процесс генерации виртуальных библиотек макролидов.

В качестве шаблонов при генерации используются 17 макролактоновых колец размером от 6-ти до 38-ми атомов (рис. 2). Для генерации структурных формул макролидов эти макролактоновые кольца условно разбиваются на фрагменты, каждый фрагмент соответствует одному «строительному блоку» (рис. 3).

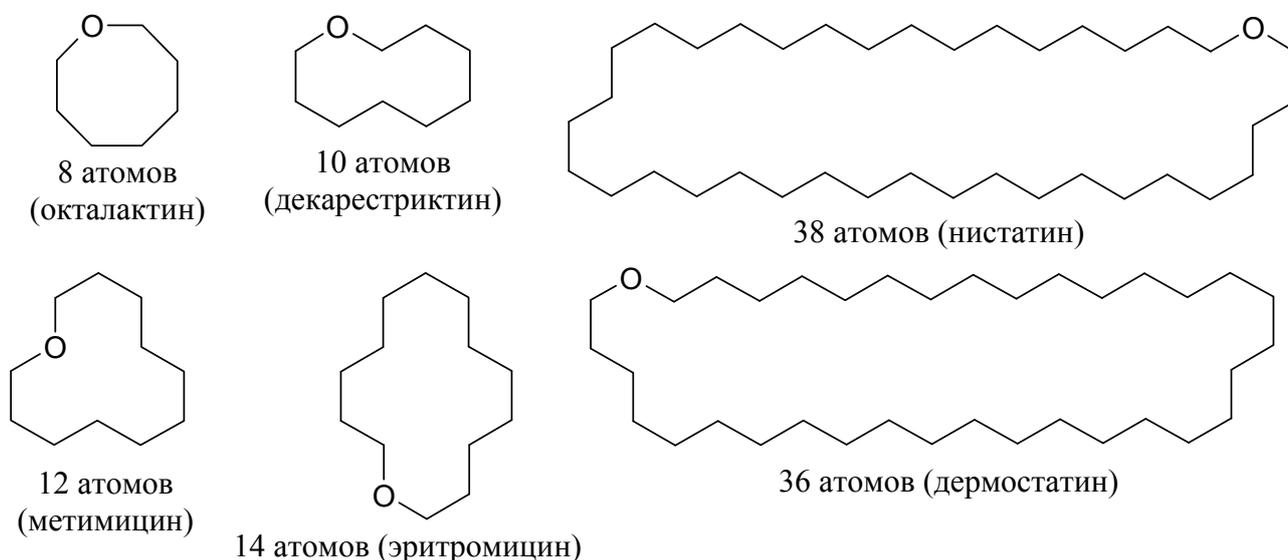


Рис. 2. Примеры макролактоновых колец различного размера.

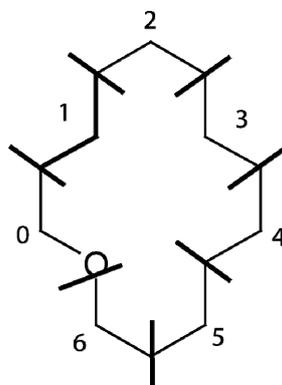


Рис. 3. Разбиение макролактонового кольца из 14-ти атомов на фрагменты.

При построении структурных формул макролидов, исходя из доменного состава синтезирующих их ферментов, учитываются следующие данные о строении и функционировании модульных PKS:

1. В составе модуля обязательно присутствуют домены AT, ACP и KS.
2. Домен AT может быть малонат- или метилмалонат-специфическим. В зависимости от типа домена в молекулу макролида добавляется ацетат или пропионат (рис. 4).

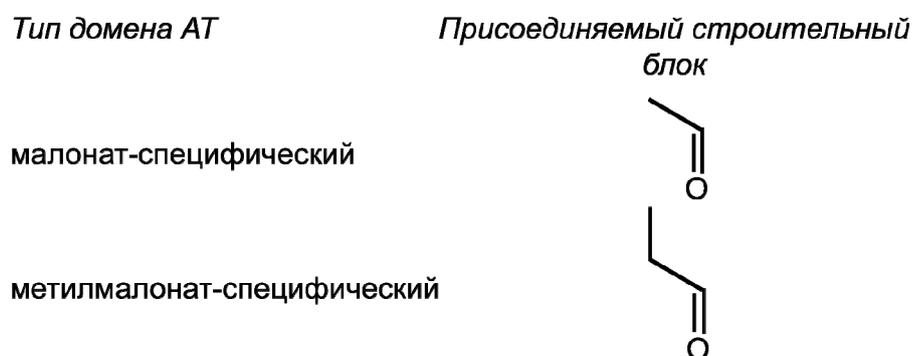


Рис. 4. Типы присоединяемых строительных блоков.

3. Различные сочетания дополнительных доменов (KR, DH, ER) обуславливают разную степень восстановления строительного блока, который был добавлен предыдущим модулем в поликетидную цепь (рис. 5). Строительный блок, добавленный последним модулем, не подвергается дальнейшему восстановлению.

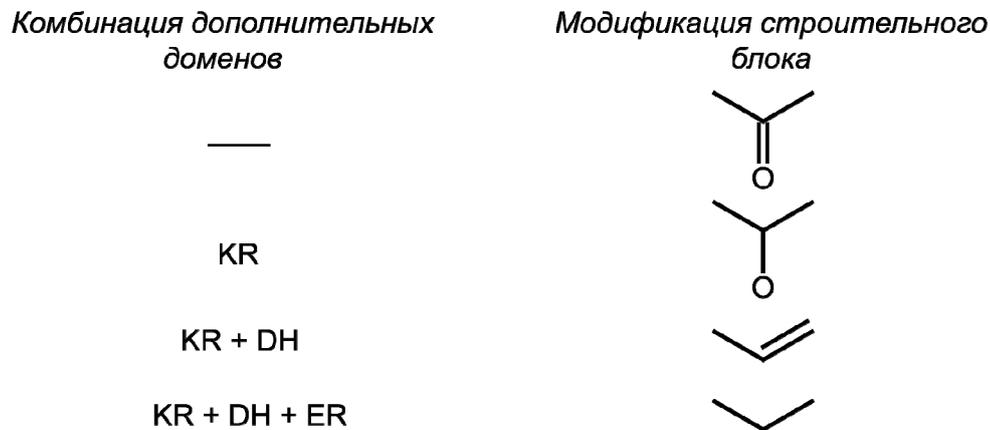


Рис. 5. Типы модификаций строительных блоков.

4. Первый модуль всегда содержит один дополнительный домен – KR, набор дополнительных доменов не может изменяться в процессе генерации. Это связано с тем, что в результате работы KR домена первого модуля образуется гидроксильная группа, которая необходима для замыкания макролактонового кольца.

Таким образом, для генерации виртуальных комбинаторных библиотек макролидов используется 8 типов фрагментов (рис. 6), строение каждого фрагмента определяется доменным составом двух соседних модулей. Генерация виртуальных библиотек производится путем перебора всех возможных комбинаций фрагментов. Фрагмент в нулевой позиции (рис. 3) не изменяется.

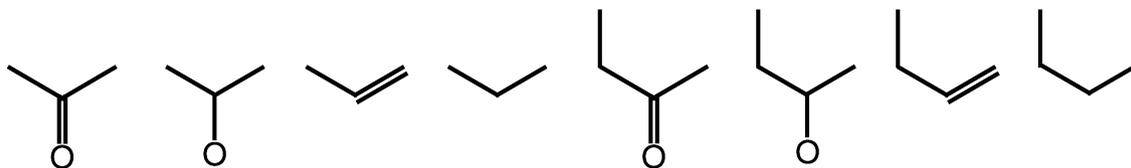


Рис. 6. Фрагменты, используемые для генерации виртуальных библиотек макролидов.

Перед началом процесса генерации в программе производится подсчет числа структур, которые могут быть сгенерированы при заданных параметрах, по формуле:

$$N = k_1 \cdot k_2 \cdot k_3 \cdot \dots \cdot k_{n-1} \cdot a$$

где n – число модулей в PKS; k_i – число возможных фрагментов в i -й позиции (в общем случае $k_i = 8$, при наложении ограничений на структуру молекулы – пост-поликетидные модификации и т.д. – это число уменьшается); a – число возможных фрагментов в последней позиции (1 или 2).

В выходные SDF-файлы записываются структуры макролидов и описания PKS, необходимых для их синтеза. Описание PKS записывается в поле “<modules>” в следующем виде:

```
> <modules>
[Loading:Propionate]
[KS-ATmal-**-**-KR-ACP][KS-ATmal-DH-ER-KR-ACP][KS-ATmal-DH-ER-KR-ACP]
[KS-ATmal-DH-ER-KR-ACP][KS-ATmal-DH-ER-KR-ACP][KS-ATmal-DH-ER-KR-ACP],
```

где доменный состав отдельных модулей заключен в квадратные скобки.

База данных по структурным формулам макролидов и соответствующим им биологическим активностям

Для пополнения обучающей выборки программы PASS нами была создана база данных (БД) Macrolides, содержащая структурные формулы макролидов и известные для них виды биологической активности. БД Macrolides имеет стандартную структуру, используемую для создания обучающей выборки программы PASS, и содержит следующие поля: ID (идентификационный номер), chem.name (химическое название соединения), structure (структурная формула соединения), class (список видов биологической активности соединения), lit.ref (ссылки на источник информации), и др.

Созданная нами БД Macrolides содержит 1114 структур макролидов, проявляющих 247 видов биологической активности. В базу входят макролиды с размером макролактонового кольца от 8 до 48 атомов. Число включенных в БД соединений с различным размером макролактонового кольца варьирует от 1 (25, 42, 44, 48 атомов в кольце) до 259 структур (16 атомов в кольце). Количество включенных в БД структур с различными размерами макролактонового кольца коррелирует с данными S. Omura о числе известных макролидов соответствующего размера [Omura S., 2002].

Прогноз биологической активности макролидов

Оценка точности прогноза биологической активности макролидов для различных обучающих выборок. В обучающей выборке программы PASS содержатся данные о биологической активности соединений, в то время как используемое в литературе понятие «макролид» характеризует структурные особенности вещества. Наиболее близким к понятию «макролид» термином, используемым при описании биологической активности соединений в программе PASS, является «Антибиотик макролидного ряда» (Antibiotic Macrolide-like). Однако не все макролиды являются антибиотиками, и, следовательно, не каждому макролиду в обучающей выборке приписана эта активность. В момент начала нашей работы в обучающей выборке PASS содержалось 216 структур с активностью «Антибиотик макролидного ряда». Это значительно меньше числа известных макролидов, которое в 2002 году составляло более 2000 структур [Omura S., 2002].

Для прогноза биологической активности макролидов нами была создана база данных Macrolides, состоящая только из макролидов. Было проведено сравнение точности прогноза проявляемых макролидами видов биологической активности при использовании различных обучающих выборок:

- 1) Исходная обучающая выборка программы PASS, содержит 216 соединений с активностью «Антибиотик макролидного ряда».
- 2) Обучающая выборка, состоящая только из макролидов (получена из базы данных Macrolides).
- 3) Объединенная обучающая выборка, состоящая из исходной обучающей выборки программы PASS и базы данных Macrolides.

При обучении программы PASS на выборке, состоящей только из макролидов, возможно предсказание 106 видов активности (в соответствии с алгоритмом отбора прогнозируемых видов активности в программе PASS, см. «Материалы и методы»). В таблице 1 приведены сведения об инвариантной точности прогноза (IAP) для некоторых видов активности, а также среднее значение IAP по 106 видам активности.

Точность прогноза различных видов биологической активности меняется при использовании различных обучающих выборок (см. таблицу 1). Для 67 видов активности точность прогноза увеличивается при добавлении макролидов к исходной обучающей выборке по сравнению с исходной выборкой, однако в 27 случаях точность уменьшается. Прогноз 12-ти видов активности невозможен при использовании только исходной обучающей выборки в соответствии с ограничениями по отбору прогнозируемых активностей, имеющимися в программе PASS. Это такие виды активности как «Стимулятор деполимеризации актина» (Actin depolymerization stimulant), «Ингибитор полимеризации актина» (Actin polymerization inhibitor) и др.

Таблица 1. Инвариантная точность прогноза (IAP) видов биологической активности, проявляемых макролидами, при использовании различных обучающих выборок.

Название активности	Исходная об. выб.		Только макролиды		Исходная об. выб. + макролиды	
	N	IAP,%	N	IAP,%	N	IAP,%
Actin depolymerization stimulant	-	-	4	80,58	4	98,57
Actin polymerization inhibitor	-	-	4	99,18	4	99,99
Angiogenesis inhibitor	570	85,99	14	72,51	584	85,98
Antiacne	306	94,07	8	90,93	314	94,11
Antiarthritic	3015	82,77	20	83,29	3035	82,63
Antibacterial	4958	91,65	612	92,66	5570	91,90
Antibiotic	3333	95,53	587	91,83	3920	95,69
Antibiotic Macrolide-like	216	99,62	504	91,13	720	99,63
....
Tubulin antagonist	111	91,13	10	92,37	121	92,08
Tyrosine kinase inhibitor	1632	96,69	5	85,29	1637	96,65
Среднее значение		88,0		89,8		89,4

N – количество соединений с экспериментально подтвержденной активностью данного вида в обучающей выборке PASS;

IAP, % – инвариантная точность прогноза по скользящему контролю с исключением по одному для обучающей выборки;

« - » – данный вид активности не прогнозируется с помощью исходной обучающей выборки.

В целом, средняя точность прогноза при добавлении макролидов к исходной обучающей выборке (89,4%) несколько увеличивается по сравнению с исходной обучающей выборкой (88%). Средняя IAP при обучении только на макролидах (89,8%) выше, чем при использовании исходной обучающей выборки (88%), однако незначительно уменьшается при добавлении макролидов к исходной обучающей выборке (89,4%).

Разница в средней точности прогноза при обучении только на макролидах и при использовании объединенной обучающей выборки незначительна, при этом объединенная выборка позволяет прогнозировать значительно большее число видов биологической активности (2806 видов активности). Поэтому при дальнейшем прогнозировании биологической активности мы использовали наиболее полную обучающую выборку, состоящую из исходной выборки PASS и добавленных к ней макролидов.

Оценка точности прогноза биологической активности макролидов на тестовой выборке. Для оценки результатов применения предложенного нами подхода для поиска новых биологически активных макролидов необходимо синтезировать и протестировать достаточно большое количество соединений. В то же время, методами

генной инженерии можно получить лишь небольшое количество бактериальных штаммов, которые будут продуцировать отобранные соединения. Это весьма длительный процесс, причем экспериментальных данных будет заведомо недостаточно для статистически значимого заключения о пригодности предложенного метода. В связи с этим, валидация метода возможна только на известных из литературы экспериментальных данных.

Нами была проведена оценка применимости программы PASS для предсказания биологической активности макролидов на независимой тестовой выборке из 242 природных макролидов. Выборка была составлена на основе базы данных Dictionary of Natural Products (DNP). В тестовую выборку включены вещества, которые относятся к группе «Macrolide polyketides» и «Polyenes» по классификации этой базы данных.

Результаты предсказания биологической активности для этой выборки представлены в таблице 2.

Таблица 2. Результаты прогноза биологической активности соединений из тестовой выборки (242 природных макролида).

Название активности	N	IAP	N (PASS)	IAP (PASS)
Антибактериальная	62	0,7456	5570	0,9190
Антибиотик	61	0,6335	3920	0,9569
Антибиотик макролидного ряда	58	0,6963	720	0,9963
Токсическое действие	50	0,8492	1626	0,8182
Цитотоксическая	49	0,8636	293	0,9415
Противогрибковая	43	0,6427	1634	0,8839
Противопаразитарная	36	0,9981	548	0,8939
Противоопухолевая	36	0,9260	8886	0,8440
Противоглистная	15	0,9430	422	0,8966
Иммуносупрессорная	12	0,4844	1081	0,8116
Противовирусная	10	0,8043	3968	0,8767
Инсектицидная	6	0,8482	127	0,9232
....
Среднее значение		0,7553		0,8942

N – количество соединений с экспериментально подтвержденной активностью данного вида в тестовой выборке;

IAP – инвариантная точность прогноза для тестовой выборки;

N (PASS) – количество соединений с экспериментально подтвержденной активностью данного вида в обучающей выборке PASS;

IAP (PASS) – инвариантная точность прогноза по скользящему контролю с исключением по одному для обучающей выборки.

Таблица 2 содержит данные о количестве соединений в тестовой выборке и обучающей выборке PASS, проявляющих тот или иной вид активности в эксперименте, а также оценку точности прогноза (IAP) для каждой активности. Всего с помощью

программы PASS возможен прогноз 41 вида активности, известного для соединений из тестовой выборки. В автореферате представлены только те виды активности, для которых в тестовой выборке содержится 6 и более соединений (полностью таблица приведена в диссертации). Вычисление IAP для обучающей выборки проводилось в режиме скользящего контроля с исключением по одному. В обучающей выборке PASS содержится 33 структуры из тестовой выборки, однако при выполнении прогноза для каждого из этих соединений соответствующие данные исключались из обучающей выборки.

Для обучающей выборки все виды активности прогнозируются с высокой точностью: от 77,5% (антагонист интерлейкина 1) до 99,7% (стимулятор протеинкиназы C). Как видно из таблицы 2, для тестовой выборки виды активности, для которых имеется достаточно большое количество соединений, прогнозируются с приемлемой точностью: от 63,4% (антибиотик) до 99,8% (противопаразитарная). Исключением является иммуносупрессорная активность (48,4%). Оценка точности прогноза для активностей, представленных небольшим количеством соединений (данные не приведены в таблице 2), не является статистически достоверной. В этих случаях IAP может варьировать в широких пределах: от 0,8% (ингибитор синтеза белка) до 100% (ингибитор синтеза холестерина).

Средняя точность прогноза по всем прогнозируемым видам активности для тестовой выборки, состоящей только из макролидов, составляет 75,5%. Средняя точность прогноза для обучающей выборки, включающей соединения разных классов – 89,4%. Таким образом, даже для тестовой выборки, состоящей из близких по структуре соединений, PASS позволяет прогнозировать различия в спектре активности с приемлемой точностью.

Оценка точности прогноза биологической активности полусинтетических аналогов эритромицина. Макролиды природного происхождения часто модифицируют химически для улучшения их фармакодинамических и фармакокинетических характеристик. На основе эритромицина было получено множество полусинтетических производных (рокситромицин, кларитромицин, азитромицин и др.). Нами была проведена оценка применимости программы PASS для предсказания биологической активности полусинтетических аналогов эритромицина на независимой тестовой выборке, содержащей 612 соединений из базы данных MDDR.

В таблице 3 представлены данные о количестве соединений в тестовой выборке, проявляющих тот или иной вид активности в эксперименте, а также оценки точности прогноза (IAP) для каждой активности. Всего с помощью программы PASS возможен прогноз 27 видов активности, известных для соединений из тестовой выборки. В

автореферате представлены виды активности, для которых в тестовой выборке содержится 5 и более соединений (полностью таблица приведена в диссертации).

Таблица 3. Результаты прогноза биологической активности соединений из тестовой выборки (612 полусинтетических аналогов эритромицина).

Название активности	N	IAP
Антибактериальная	518	0,6771
Антибиотик	504	0,6101
Антибиотик макролидного ряда	341	0,6539
Противопротозойная	55	0,8641
Противоартритная	23	0,9738
Прокинетик	21	0,9895
Антагонист рилизинг-фактора лютеинизирующего гормона	19	1,0
Агонист рецепторов мотилина	16	0,9850
Противоопухолевая	15	0,9655
Антибиотик хинолонового ряда	13	0,9787
Противомикобактериальная	13	0,9935
Противоастматическая	13	0,5811
Вещество для лечения акне	11	0,9624
Ингибитор выработки цитокинов	6	0,4757
Вещество для лечения рака простаты	5	0,9898
....
Среднее значение		0,8288

N – количество соединений с экспериментально подтвержденной активностью данного вида в тестовой выборке;

IAP – инвариантная точность прогноза для тестовой выборки.

Для тестовой выборки из полусинтетических аналогов эритромицина точность прогноза различных видов активности колеблется от 61% (антибиотик) до 100% (антагонист рилизинг-фактора лютеинизирующего гормона). В целом, для точности прогноза различных видов активности веществ из данной тестовой выборки наблюдается та же закономерность, что и для тестовой выборки из 242 природных макролидов (см. выше). Не высокие значения IAP получены для видов активности, относящихся к воздействию соединений на иммунную систему: «Ингибитор выработки цитокинов» и «Противоастматическая» (данная тестовая выборка), и «Иммуносупрессорная» (тестовая выборка из 242 природных макролидов). Оценки точности прогноза для активностей, представленных небольшим количеством соединений, не являются статистически достоверными, и значения IAP варьируют в широких пределах.

Средняя точность прогноза по 27 прогнозируемым видам активности полусинтетических аналогов эритромицина – 82,9%, что позволяет использовать программу PASS для прогноза биологической активности аналогов эритромицина.

Генерация и скрининг виртуальной библиотеки аналогов эритромицина

Для апробации разработанного подхода мы сгенерировали виртуальную библиотеку аналогов эритромицина с помощью программы Biogenerator и провели ее компьютерный скрининг.

Генерация аналогов эритромицина была выполнена при следующем наборе параметров: пропионат в качестве стартового блока, 6 модулей расширения, гликозилирование кладинозой по 3-му атому углерода, гликозилирование десозамином по 5-му атому углерода, гидроксильное по 6-му и 12-му атомам углерода. Выбор параметров генерации обусловлен строением и механизмом синтеза эритромицина. Полученная виртуальная библиотека состоит из 3072 структур.

Для сгенерированных структур был выполнен прогноз спектров биологической активности с помощью программы PASS. Результаты прогноза были проанализированы с помощью программы PharmaExpert. Были отобраны те виды активности, вероятность проявления которых выше 70% ($P_a > 0,7$) хотя бы для одной структуры.

С вероятностью более 70% были предсказаны такие известные для эритромицина и его аналогов виды активности как «Антибактериальная», «Антибиотик», «Антибиотик макролидного ряда», «Ингибитор синтеза белка», «Ингибитор 50s субъединицы рибосомы». Помимо этого, был предсказан ряд видов активности, которые были экспериментально открыты в последние годы для эритромицина и его аналогов и не связаны с его антибактериальной активностью. В частности, такие экспериментально подтвержденные виды активности как «Антагонист HERG каналов» [Stanat S.J., 2003; Chen S.Z., 2005] и «Удлинение интервала QT» [Abu-Gharbieh E., 2004], «Субстрат CYP3A» и «Субстрат CYP3A4» [Rodrigues A.D, 1997; Kenworthy K.E., 1999], «Противопаразитарная» [Degerli K., 2003], «Гепатотоксическая» [Aronson J.K., 2006], «Вещество для лечения дискинезии» и «Стимулятор моторики желудочно-кишечного тракта» [Peeters T.L., 1993; Doherty W.L., 2003] и др.

Для некоторых прогнозируемых видов активности в доступной литературе не было найдено экспериментального подтверждения, в частности для таких видов активности как «Противопаразитарная», «Ингибитор синтеза холестерина» и др. Поскольку такие виды активности были предсказаны только для некоторых структур из сгенерированной комбинаторной библиотеки, возможно они специфичны для этих соединений и никогда не проверялись для известных аналогов эритромицина.

Далее мы провели компьютерный скрининг библиотеки аналогов эритромицина. Эритромицин сравнительно хорошо переносится пациентами, наиболее серьезным побочным эффектом, наблюдаемым при приеме всех форм эритромицина, является холестаза [Aronson J.K., 2006]. Прием эритромицина может приводить к повышению уровня сывороточных трансаминаз и развитию холестатического гепатита. Механизм

развития повреждения печени в результате приема эритромицина не ясен: показано, что эритромицин не ингибирует транспорт желчных кислот (таурохолевой кислоты) [Kostrubsky V.E., 2003]. Был проведен поиск аналогов эритромицина, для которых вероятность проявления гепатотоксической активности наименьшая.

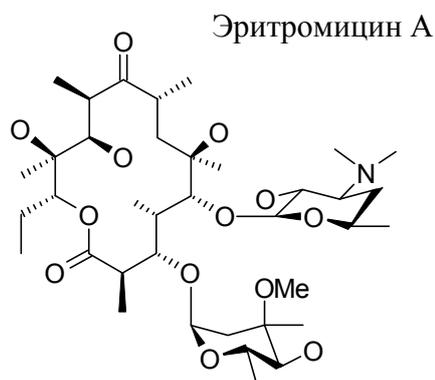
Использовалась сгенерированная ранее библиотека из 3072 аналогов эритромицина. Отбор соединений проводили при значениях $P_a > 0,8$ для активности «Антибактериальная» и $P_a < 0,6$ для активности «Гепатотоксическая». Инвариантная точность прогноза гепатотоксической активности составляет 75,9%, антибактериальной – 91,9%.

В результате было отобрано 17 аналогов эритромицина, удовлетворяющих заданным критериям. На рисунке 7 приведены значения P_a для антибактериальной и гепатотоксической активностей эритромицина и его аналогов, а также отношение значений P_a для этих видов активности.



Рис. 7. Результаты прогноза антибактериальной и гепатотоксической активностей для эритромицина (эр.) и его 17-ти аналогов, отобранных при генерации и скрининге виртуальной библиотеки. P_a (Гп) – вероятность наличия гепатотоксической активности; P_a (Аб) – вероятность наличия антибактериальной активности; P_a (Аб) / P_a (Гп) – отношение P_a (Аб) к P_a (Гп).

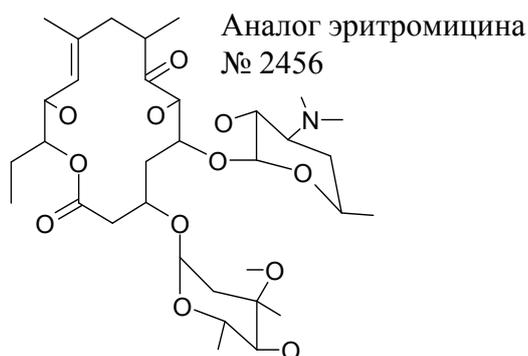
Из рисунка 7 видно, что наименьшее значение P_a для гепатотоксической активности при наибольшем значении P_a для антибактериальной активности прогнозируется для молекулы № 2456. В соответствии с результатами предсказания, вероятность проявления гепатотоксичности для молекулы №2456 составляет 56,5%, что существенно ниже, чем у эритромицина (94,1%). Таким образом, при наличии возможности выбора аналогов эритромицина, наиболее перспективным является аналог №2456 (рис. 8). Необходимая для синтеза этой молекулы информация о доменном составе PKS (рис. 8) получена с помощью программы Biogenerator и является основой для конструирования соответствующего микроорганизма-продуцента.



[Loading: Propionate]
 [KS-ATmet-**-**-KR-ACP]
 [KS-ATmet-**-**-KR-ACP]
 [KS-ATmet-**-**-**-ACP]
 [KS-ATmet-DH-ER-KR-ACP]
 [KS-ATmet-**-**-KR-ACP]
 [KS-ATmet-**-**-KR-ACP]

PASS ACTIVITY SPECTRUM at Pa > 0.8

Pa	Pi	
0.991	0.001	Antibiotic Macrolide-like
0.990	0.003	Cytochrome P450 inhibitor
0.982	0.000	Antibacterial
0.975	0.000	Antibiotic
0.969	0.001	Mucolytic
0.968	0.002	Protein 50S ribosomal subunit inhibitor
0.961	0.000	Transcription factor NF kappa B inhibitor
0.957	0.003	Antiprotozoal (Toxoplasma)
0.955	0.003	Protein synthesis inhibitor
0.951	0.003	HERG channel antagonist
0.943	0.005	QT interval prolongation
0.937	0.000	Antirickettsial
0.941	0.004	Hepatotoxic
0.932	0.001	Expectorant
0.922	0.002	Transcription factor inhibitor
0.870	0.003	Gastrointestinal motility stimulant
0.853	0.002	Prokinetic
0.844	0.014	Toxic
0.819	0.005	Antidyskinetic



PASS ACTIVITY SPECTRUM at Pa > 0.8

Pa	Pi	
0.904	0.007	Antiprotozoal (Toxoplasma)
0.888	0.001	Transcription factor NF kappa B inhibitor
0.885	0.009	Toxic
0.869	0.001	Antibacterial
0.868	0.004	Transcription factor inhibitor
0.845	0.001	Antibiotic
0.839	0.006	HERG channel antagonist
...		
0.565	0.109	Hepatotoxic

[Loading: Propionate]
 [KS-ATmal-**-**-KR-ACP]
 [KS-ATmet-DH-**-KR-ACP]
 [KS-ATmet-DH-ER-KR-ACP]
 [KS-ATmal-**-**-**-ACP]
 [KS-ATmal-**-**-KR-ACP]
 [KS-ATmal-**-**-KR-ACP]

Рис. 8. Структурные формулы и прогнозируемые PASS виды биологической активности для эритромицина А и его аналога №2456. Состав модулей PKS типа I, необходимой для синтеза аналога №2456, получен с помощью программы Biogenerator. ATmal – малонил-специфический ацилтрансферазный домен, ATmet – метилмалонил-специфический ацилтрансферазный домен. Остальные пояснения – см. в тексте.

На рисунке 8 курсивом выделены домены, которые отличаются в DEBS эритромицина и гипотетической поликетидсинтазе, необходимой для синтеза аналога №2456. Как видно из рисунка, доменный состав модулей в ферментативных системах, необходимых для биосинтеза двух структур, значительно отличается. Сконструировать поликетидсинтазу, способную синтезировать аналог №2456, из DEBS можно путем изменения специфичности АТ доменов или числа дополнительных доменов в каждом

из 6-ти модулей DEBS, или замены целых модулей на модули с другим доменным составом. В литературе описаны успешные эксперименты по внесению множественных изменений в DEBS, показано что при внесении 3-х изменений фермент остается функциональным [McDaniel, 1999], а также эксперименты по комбинированию в единую ферментативную систему белков, составляющих PKS эритромицина, пикромицина и олеандомицина [Tang, 2000].

Из 17-ти отобранных аналогов эритромицина, минимальное число изменений в доменном составе DEBS необходимо для конструирования ферментативной системы, способной синтезировать аналог №2310. Это изменение субстратной специфичности АТ домена в 1-м и 6-м модулях, введение дополнительного домена DH во втором модуле, замена домена АТ и удаление дополнительных доменов в 4-м модуле:

DEBS эритромицина

[Loading:Propionate]
[KS-ATmet-**-**-KR-ACP]
[KS-ATmet-**-**-KR-ACP]
[KS-ATmet-**-**-**-ACP]
[KS-ATmet-DH-ER-KR-ACP]
[KS-ATmet-**-**-KR-ACP]
[KS-ATmet-**-**-KR-ACP]

Аналог №2310

[Loading:Propionate]
[KS-*ATmal*-**-**-KR-ACP]
[KS-ATmet-*DH*-**-KR-ACP]
[KS-ATmet-**-**-**-ACP]
[KS-*ATmal*-**-**-**-ACP]
[KS-ATmet-**-**-KR-ACP]
[KS-*ATmal*-**-**-KR-ACP]

ЗАКЛЮЧЕНИЕ

В диссертационной работе предложен метод конструирования *in silico* доменных последовательностей ферментативных систем, обеспечивающих синтез макролидов с заданным профилем биологической активности. Разработана компьютерная программа Biogenerator, которая позволяет генерировать виртуальные комбинаторные библиотеки структур макролидов и последовательности доменов PKS, необходимые для их биосинтеза. Показано, что применение программы PASS для прогноза биологической активности и программы PharmaExpert для анализа результатов прогноза, позволяет отбирать среди сгенерированных структур макролиды с требуемым профилем биологической активности.

С целью оптимизации расчетов в случаях, когда количество сгенерированных структур очень велико, нами была разработана компьютерная программа BioGenPharm. Программа BioGenPharm интегрирует функции программ Biogenerator, PASS и PharmaExpert, и позволяет генерировать и отбирать виртуальные структуры поликетидов без записи промежуточных результатов.

Предложенный подход может быть использован для поиска новых соединений, синтезируемых другими видами многомодульных ферментативных систем, например нерибосомальными пептидными синтазами.

ВЫВОДЫ

1. Созданная программа Biogenerator для генерации виртуальных библиотек структурных формул макролидов и описаний синтезирующих их ферментативных систем, достаточных для генно-инженерной разработки продуцентов соответствующих макролидов, обеспечивает исчерпывающий перебор потенциально возможных структур.
2. Показано, что добавление к стандартной обучающей выборке программы PASS информации из созданной нами базы данных по макролидам позволяет прогнозировать биологическую активность макролидов со средней точностью 89,4%.
3. Используемый метод прогноза биологической активности предсказывает биологическую активность макролидов со средней точностью 75,5% для тестовой выборки, состоящей из природных макролидов, а для тестовой выборки, состоящей из полусинтетических аналогов эритромицина, со средней точностью 82,9%.
4. На примере генерации и скрининга виртуальной библиотеки аналогов эритромицина показано, что разработанный метод позволяет отбирать среди сгенерированных структур вещества с требуемым профилем биологической активности и определять доменный состав ферментативных систем, необходимый для их биосинтеза.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Zotchev S.B., Stepanchikova A.V., Sergeyko A.P., Sobolev B.N., Filimonov D.A., Poroikov V.V. Rational design of macrolides by virtual screening of combinatorial libraries generated through in silico manipulation of polyketide synthases. // *J Med Chem.* – 2006. Vol. 49. № 6. – P. 2077-2087.
2. Сергейко А.П., Степанчикова А.В., Соболев Б.Н., Зотчев С.Б., Лагунин А.А., Филимонов Д.А., Поройков В.В.. Компьютерное конструирование поликетидов с заданными свойствами. // *Биомедицинская химия.* – 2007. – Т. 53. № 5. – С. 522-531. (Sergeyko A.P., Stepanchikova A.V., Sobolev B.N., Zotchev S.B., Lagunin A.A., Filimonov D.A., Poroikov V.V. Computer-aided design of polyketides with the required properties. // *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry* – 2008. – Vol. 2. № 1. – P. 55-62).
3. Зотчев С.Б., Степанчикова А.В., Сергейко А.П., Соболев Б.Н., Филимонов Д.А., Поройков В.В. Генерация in silico и скрининг виртуальных библиотек макролидов. // *Московская международная конференция “Биотехнология и медицина”.* Материалы конференции. – Москва. – 2006. – С. 44.

4. Сергейко А.П., Степанчикова А.В., Соболев Б.Н., Зотчев С.Б., Поройков В.В. Компьютерное проектирование ферментативных систем, обеспечивающих синтез макролидов с заданными свойствами. // Материалы XIV международной конференции и дискуссионного научного клуба «Новые информационные технологии в медицине, биологии, фармакологии и экологии». Приложение научно-практического журнала «Открытое образование» – 2006. – №3. – С. 78-79.
5. Сергейко А.П., Степанчикова А.В., Соболев Б.Н., Филимонов Д.А., Зотчев С.Б., Поройков В.В. Biogenpharm – программа для генерации виртуальных библиотек макролидов и отбора структур с заданными свойствами. // Материалы IV Московского международного конгресса «Биотехнология: состояние и перспективы развития». – Москва. –2007. – С. 409.
6. Sergeyko A., Stepanchikova A., Sobolev B., Zotchev S., Filimonov D., Lagunin A., Poroikov V. Computer – aided design of polyketides with the required properties. // Abstr. 4th Internat. Symp. «Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (СМТПИ-2007)». – Moscow (Russia). – 2007. – P. 144.
7. Зотчев С.Б., Соболев Б.Н., Степанчикова А.В., Сергейко А.П., Поройков В.В. Свидетельство об официальной регистрации программы для ЭВМ Biogenerator № 2007610742 от 15 февраля 2007 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.
8. Зотчев С.Б., Соболев Б.Н., Степанчикова А.В., Сергейко А.П., Филимонов Д.А., Лагунин А.А., Глоризова Т.А., Поройков В.В. Свидетельство об официальной регистрации программы для ЭВМ Biogenpharm № 2007610741 от 15 февраля 2007 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.

Благодарности. Работа выполнена при финансовой поддержке Норвежского исследовательского фонда (грант № 165190/V40), фирм SINVENT AS и Biosergen AS (Норвегия).

СПИСОК ИСПОЛЬЗОВАННЫХ СОКРАЩЕНИЙ

БД – база данных

ACP – acyl carrier protein (ацилпереносящий белок)

AT – acyltransferase (ацилтрансфераза)

DH – dehydratase (дегидратаза)

ER – enoyl reductase (енол-редуктаза)

IAP – Invariant Accuracy of Prediction (инвариантная точность прогноза)

KR – ketoreductase (кеторедуктаза)

KS – β -ketoacylsynthase (β -кетосинтаза)

MDDR – MDL Drug Data Report

NCE – New Chemical Entity

PASS – Prediction of Activity Spectra for Substances

PKS – polyketide synthase (поликетидсинтаза)

QSAR – Quantitative Structure-Activity Relationship (количественная взаимосвязь «структура – активность»)

SDF-файл –structure-data file

TE – thioesterase (тиоэстераза)